Accepted author manuscript, to appear/in press:
International Journal of Geographical Information Science (IJGIS)
DOI: https://doi.org/10.1080/13658816.2019.1641715

Extracting activity patterns from taxi trajectory data: A two-layer framework using spatio-temporal clustering, Bayesian probability, and Monte Carlo simulation

Shuhui Gong[a,d], John Cartlidge[b], Ruibin Bai[c], Yang Yue[d], Qingquan Li[d] and Guoping Qiu[e]

[a]International Doctoral Innovation Centre, University of Nottingham Ningbo China, Ningbo China
[b]Department of Computer Science, University of Bristol, Bristol UK
[c]Department of Computer Science, University of Nottingham Ningbo China, Ningbo China
[d]Department of Urban Informatics, School of Architecture and Urban Planning, Shenzhen University, Shenzhen China
[e]School of Computer Science, University of Nottingham, Nottingham UK

ARTICLE HISTORY
Compiled July 7, 2019

ABSTRACT
Global positioning system (GPS) data generated from taxi trips is a valuable source of information that offers an insight into travel behaviours of urban populations with high spatio-temporal resolution. However, in its raw form, GPS taxi data does not offer information on the purpose (or intended activity) of travel. In this context, to enhance the utility of taxi GPS data sets, we propose a two-layer framework to identify the related activities of each taxi trip automatically and estimate the return trips and successive activities after the trip, by using geographic point-of-interest (POI) data and a combination of spatio-temporal clustering, Bayesian inference, and Monte Carlo simulation. Two million taxi trips in New York, the United States of America, and ten million taxi trips in Shenzhen, China, are used as inputs for the two-layer framework. To validate each layer of the framework, we collect 6,003 trip diaries in New York and 712 questionnaire surveys in Shenzhen. The results show that the first layer of the framework performs better than comparable methods published in the literature, while the second layer has high accuracy when inferring return trips.

KEYWORDS
Spatio-temporal clustering; Bayesian Probabilities; Monte Carlo simulation; Travel behaviours

1. Introduction

Daily activity analysis is of great significance for urban planning (Jones 1990, Beecham *et al.* 2014). A clear understanding of people's daily activities can help the government to plan the urban infrastructure more rationally. Previously, GPS data has been used for analysing different activities, such as geographical model calibration (Gong *et al.* 2017), discovering shopping patterns (Gong *et al.* 2016), and modelling demand for

CONTACT Shuhui Gong. Email: shuhui.gong@nottingham.edu.cn

point-of-interest (POI) (Liu *et al.* 2017). However, it is challenging to connect people's activities with their travel routines (i.e. identify activities from trajectory data). Previous work called this challenge "activity inference" (Gong *et al.* 2016). Although GPS taxi data includes accurate individual locations, the technical question of how to discover the correlations describing arrivals and tracing trip purpose from such data remains difficult to answer. The given data may be rich, but the activity information is sparse (Gong *et al.* 2016, Wang *et al.* 2017).

Previous studies proposed methods to estimate trip purpose based on taxi trajectory data. Yue *et al.* (2012) defined a simple "buffer" radius, based on anchor stores, to reflect the catchment of a shopping centre neighbourhood such that all taxi drop-off points (DOPs) located within the buffer zone are assumed to denote the beginning of a shopping trip. Xie *et al.* (2009) used a similar method to identify a trip's purpose (or related *activity*), as being associated with a POI nearest to a taxi's DOP; Huang *et al.* (2010) set up a model defining the attractiveness of the POI, calculated by the POI size; Gong *et al.* (2016) extended Huang's work by using the Bayesian probability inference on two factors—the opening time of each POI and the distance between POI and DOP.

We find gaps that limit previous research in this area. First, trip purpose is inferred using only DOP locations and POI opening times. However, it has been shown that taxi origin or pick-up point (PUP) and temporal information about a taxi's destination (such as drop-off time), or DOP, are closely related to trip purpose (Wu *et al.* 2014). Therefore, the models can be improved by including this additional information. Second, previous research tends to estimate activities using the simplifying assumptions of linearity. For example, Furletti *et al.* (2013) proposed a method based on a gravity model, which assumes that the size of a shopping centre and trip distance are linearly related to trip purpose. Finally, often, there are deterministic assumptions, such that the activity is the nearest POI to the DOP. However, such determinism cannot account for two trips with identical DOPs that are aimed for different activities. A probabilistic model should be employed to capture this non-determinism.

To close the identified research gaps, we develop a two-layer framework to connect passengers' trips with activities and, after the trip, infer their return trips and successor activities. In the first layer, we develop an *activity inference model* (AIM) to label trip activities. In the second layer, we develop a *pairing journey model* (PJM) to identify successor journeys in the data. The framework employs exponential distance decay functions, K-means clustering, and Bayesian inference to infer the intended activities of a DOP located near to multiple POIs, while Monte Carlo simulation is used to model individual non-deterministic behaviours. The paper is organised as follows. In Section 2, we review related works. We introduce our two-layer framework in Section 3. In Section 4, we give a detailed explanation on the first layer—the activity inference model (AIM). Two studies in New York and Shenzhen are conducted using the proposed AIM, with results validated using the trip diaries in New York, and questionnaires in Shenzhen. We also compare the accuracy between AIM and three other methods in the existing literature: Yue *et al.* (2012), Furletti *et al.* (2013), and Gong *et al.* (2016). In Section 5, we introduce the second layer—the pairing journeys model (PJM)—to infer return trips after different activities. PJM results are validated using taxi data and questionnaires in Shenzhen. Finally, Section 6 concludes this study.

2. Literature review

2.1. *Methods for inferring geographical activity*

Vehicle GPS data is crucial for intelligent transportation systems (ITS). In order to understand human mobility and provide insights for urban planning, previous studies used GPS data to uncover spatio-temporal trip routines with related activities. In this direction, three representative related methods used are the simple buffer radius (Yue *et al.* 2012), Furletti's model (Furletti *et al.* 2013), and Gong's model (Gong *et al.* 2016).

The simple buffer radius helps in the understanding of shopping behaviours using taxi trips. Therefore, in this case, it is essential to filter the data to collect shopping trips only. Yue et al. took a simple approach, assuming that all taxi DOPs near major shopping centres denote the beginning of shopping trips (Yue *et al.* 2012). To define this locale, a "buffer radius" around the shopping centre was used, with a value of 500 m.

Furletti's model aims to infer trip purpose and uncover trip patterns, thereby overcoming the challenge of understanding why people travel based on GPS trajectory data. To understand the relationship between human mobility and their activities, Furletti et al. proposed a model, based on gravity model, to infer trip purpose, assuming that people's activities are closely related to DOP and the size of POI, which is the location of the activity (Furletti *et al.* 2013). Particularly, when the walking distance is short, or the POI is large, there is a high probability that the trip is aimed at reaching the POI. The study conducted a real case study of car trajectories, manually annotated by users with their activities.

Focusing on a similar target as that of Furletti, Gong et al. used Bayesian probabilities to propose a model to infer trip purpose, considering that people's trip purposes are related to the walking distance between the DOP and POI (Gong *et al.* 2016). Particularly, when the walking distance is short, there is a high probability that the trip is aimed at the POI. Moreover, the drop-off time should match the POI's opening time. The work conducted a case study using taxi data in Shanghai to infer trip purpose, and data from questionnaires were used to validate the case study by comparing the proportion of each activity. The results showed that the proposed model had high accuracy when inferring trip activities.

2.2. *Clustering for travel behaviour*

The commonality between the aforementioned studies is that they infer the trip purpose using a hand-designed heuristic, or simplistic model. The knowledge that is revealed just through the human recognition approach generally presents its limitations. These models are limited, and are unable to capture the variety of human behaviours and travel patterns exhibited in a large population. To account for this wide distribution in behaviours, a data-driven method is necessary, making use of large quantities of real-world travel records for a given city.

Data-driven travel behaviour analysis is facilitated by the continuous generation of data. Researchers can use algorithms to extract rules and knowledge from a large amount of data. Clustering is an unsupervised data mining method, which has a long history in a variety of scientific fields (Jain 2008). Clustering can distinguish different groups from raw data, according to certain characteristics, without prior knowledge. Previously, clustering has been applied to travel behaviour identification: Ashbrook

and Starner (2003) applied K-means clustering and Markov model on taxi drop-off location to estimate related activities; Yue *et al.* (2009) employed trajectory clustering on trip information to estimate attractive areas in the city.

First introduced in 1967, K-means is a widely known clustering method used for partitioning a d-dimensional population into k sets (MacQueen 1967). The formula is:

$$argmin_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = argmin_S \sum_{i=1}^{k} \|S_i\| Var(S_i) \quad (1)$$

where arg means argument, and argmin denotes the points of the function at which the function values are minimized. Var is the variance of the points. $\{x_1, x_2,...,x_n\}$ is a set of n observations, each a d-dimensional real vector, and μ_i is the mean of points in S_i. The process iteratively clusters the n observations into $k \leq n$ sets, $S=\{S_1, S_2, ..., S_k\}$, such that the within-cluster sum of squares (WCSS) is minimised. K-means has been applied successfully to perform travel behaviour analysis, including travel decision-making forecasts (Griva *et al.* 2016, Han *et al.* 2014), visitors' behaviours of websites (Pallant *et al.* 2017), and customers' purchase behaviours (Hüttel *et al.* 2018).

2.3. *Model validation*

Validation is an important step in evaluating the effectiveness of an experiment or model. It is difficult to judge the effectiveness of a model without the validation process. In the field of geographic information science and travel behaviour analysis, ground truth is often used to verify the effectiveness of the model. The term 'ground truth' is borrowed from meteorology, but, in the context of this study, it refers to what actually happens in a city. While ground truth is difficult to collect, especially when it comes to personal privacy, there are ways to gather some information for the validation process.

The proportion of activities can be considered as ground truth, referring to the percentage of each activity in all trips. For example, in a city, 25% of the trips comprise shopping trips. Gong *et al.* (2016) employed the proportion of activities to validate the results of their model. They first collected the residents' travel survey data in Shanghai and, subsequently, compared the proportion of activities in survey data with the results of their model. The study showed their model's potential to generate results close to the real situation. The distribution of trips' drop-off time can be considered as another ground truth. Wu *et al.* (2014) first used an agent-based model to simulate the patterns they found and, subsequently, used the collected distribution of the drop-off time to validate the model. Moreover, the study conducted by Raux *et al.* (2011) showed that the travel time budgets could be used as the ground truth to validate the effectiveness of the model. They first collected individual travel survey data from eight European cities and, subsequently, analysed the effect of several factors on time budgets for travel and out-of-home activities. They found a difference between the travel time budgets of shopping trips and work trips. Their finding motivates us to use the distribution of travel time as ground truth to conduct the validation process.

In this study, the methods mentioned above are used to conduct validation process according to the availability of the data. In Section 4.1.2, we compare the travel time distribution extracted from New York trip diaries with AIM's results to verify the effectiveness of AIM. In Section 4.2.2, we compare the proportion of activities taken from questionnaires in Shenzhen and AIM's results to validate the effectiveness of AIM. In Section 5.2, we use travel time distribution and the distribution of time spent on shopping to validate the performance of PJM in Shenzhen. The ground truth is

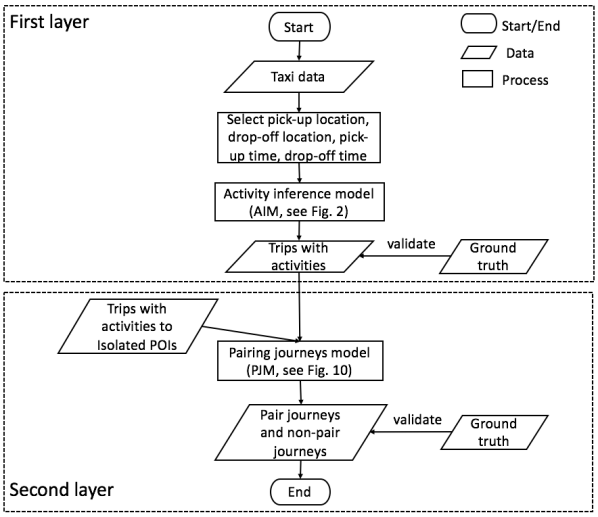


Figure 1. The proposed two-layer framework: AIM (layer one; see Fig. 2); and PJM (layer two; see Fig. 10). For ground truth validation, we use 712 questionnaires for Shenzhen and 6,003 trip diaries for New York.

collected from Shenzhen questionnaires. We also use the distribution of drop-off time, collected from Wu *et al.* (2014), to validate the results of work trips.

3. Overview of the proposed two-layer framework

The structure of the two-layer framework is shown in Fig. 1. In the first layer, we use taxi data to infer the trip purpose. Drop-off location (longitude and latitude), trip distance, and drop-off time are input variables of the AIM (see Section 4). The model clusters passengers' travel routines into different groups and indicates the related activities. The output of the AIM comprises the purposes behind the trips. To test the general performance of the AIM, we conduct two studies in both New York (Section 4.1) and Shenzhen (Section 4.2). AIM results for New York are validated with 6,003 trip diaries; AIM results for Shenzhen are validated with 712 questionnaire surveys. Finally, (Section 4.3), AIM is compared with other methods published in the literature.

In the second layer, taxi journeys and identified activities comprise the input data for the pairing journey model (PJM), which is used to estimate passengers' return trips and successor activities (see Section 5). To test the PJM, we conduct two studies using Shenzhen taxi data. First, we make use of the fact that taxi journeys with DOP near an isolated POI (IPOI) — i.e., a POI activity, such as large hospital, shop, or workplace, with no other POI located within 500 metres, as measured by network distance — can confidently be assumed to be heading to that POI, and can be tagged with an activity without the need for using the AIM (Section 5.1). This enables us to evaluate and validate the PJM in isolation. The second study directly uses the AIM results in Shenzhen to analyse their return trips (Section 5.2), enabling us to evaluate and validate the full two-layer framework. For both studies, shopping activities are validated using 712 questionnaires, while the work activity is validated by agent-based simulation's results extracted from the literature.

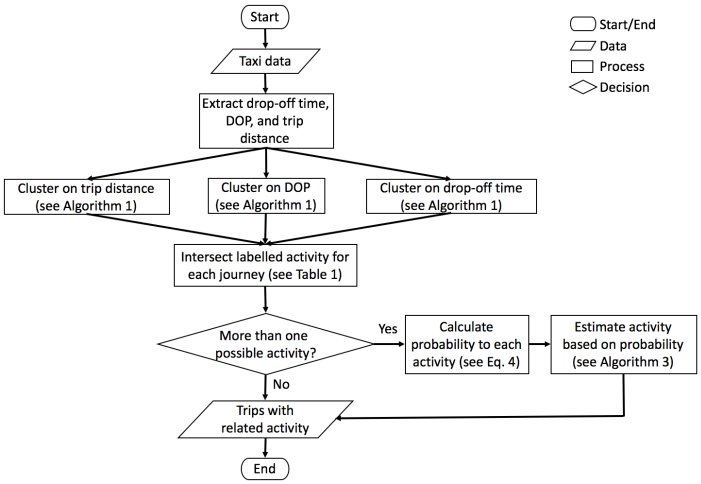
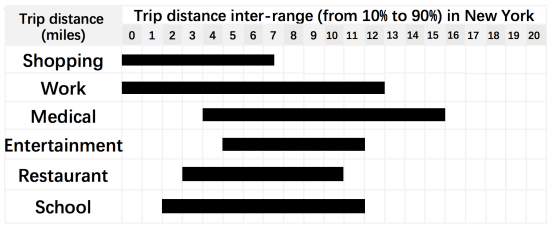


Figure 2. AIM, the process of activity inference.

4. Layer One: Activity Inference Model (AIM)

Here, we outline the first layer—the AIM process, which is used to infer the purpose of a taxi trip. The overall process is outlined in Fig. 2. First, we extract individual origin-destination (OD) taxi trips, including pick-up point PUP (location and time) and DOP (location and time). Subsequently, we consider the following three dimensions: drop-off time, walking distance (from DOP to each activity; here, we select the nearest POI as the destination if there are more than one POIs for same activity), and trip distance (calculated as the network distance, using road maps). These dimensions are chosen for the following reasons. First, activity schedules have strong time regularities, related to opening times of POIs. Second, a DOP is likely to be close to (within 500 m) an intended POI (Yue *et al.* 2012). Finally, we are interested in exploring whether there is a correlation between trip distance and an intended activity. Our aim is to infer the most likely purpose (activity) of each taxi trip. First, we perform K-means clustering independently on each of the three aforementioned dimensions. Since clustering is performed on each dimension independently, the normalisation of data is unnecessary.

(a) Trip distance distributions to three shopping IPOIs in New York.



(b) Range of trip distances for activities in New York.

Figure 3. Trip distances for New York, calculated using IPOIs.

For trip distance clustering (Algorithm 1, $C = TRIP\_DISTANCE$), the trip distance distribution is collected for trips to isolated POIs (IPOIs). A POI is considered an IPOI if there is no other POI located within 500 metres. Three IPOIs for each of the eight activities are selected, and mean travel distance distributions are recorded.

Algorithm 1 AIM clustering

Input: C, $data[x, d_x, dop_{xj}, dot_x]$ ▷ C cluster on, data: x trip; d_x trip distance; dop_{xj} DOP distance to activity j; dot_x DOT
Output: $data[x, j, b_{xj}]$ ▷ data: x trip; j activity; b_{xj} boolean 1 if x related to j, 0 otherwise

for all x **do** ▷ for each taxi trip, x
  for all j **do** ▷ for each activity, j
    $b_{xj} = 0$ ▷ Initialise activity array: set boolean zero (activity not related to trip)

if $C = TRIP\_DISTANCE$ **then**
  for k=3 to 20 **do** ▷ K-means cluster on trip distance
    clusteredData[k]=K-means(k,$data[x, d_x]$)
  K=elbow(clusteredData, 3, 20) ▷ use Elbow method (Alg. 2) to select best value of K
  for $i = 1$ to K **do** ▷ for each cluster i
    for $j = 1$ to n **do** ▷ for each activity j
      if $mean(d_i) < max(d_j)$ **then** ▷ if mean trip distance of cluster i < max trip distance of activity j
        for x in i **do** ▷ all trips in cluster i
          $b_{xj} = 1$ ▷ activity j is related to trip x in cluster i

else if $C = DROP\_OFF\_POINT$ **then**
  for k=3 to 20 **do** ▷ K-means cluster on DOP
    clusteredData[k]=K-means(k,$data[x, dop_{xj}]$)
  K=elbow(clusteredData, 3, 20) ▷ use Elbow method (Alg. 2) to select best value of K
  for $i = 1$ to K **do** ▷ for each cluster i
    for $j = 1$ to n **do** ▷ for each activity j
      if $mean(dop_{ij}) < 500$ **then** ▷ if mean distance from DOP of cluster i to activity j is < 500m
        for x in i **do** ▷ all trips in cluster i
          $b_{xj} = 1$ ▷ activity j is related to trip x in cluster i

else if $C = DROP\_OFF\_TIME$ **then**
  for k=3 to 20 **do** ▷ K-means cluster on DOT
    clusteredData[k]=K-means(k,$data[x, dot_x]$)
  K=elbow(clusteredData, 3, 20) ▷ use Elbow method (Alg. 2) to select best value of K
  for $i = 1$ to K **do** ▷ for each cluster i
    for $j = 1$ to n **do** ▷ for each activity j
      if $min(open_j) \leq mean(dot_i) \leq max(open_j)$ **then** ▷ if cluster mean DOT when activity open
        for x in i **do** ▷ all trips in cluster i
          if $min(open_j) \leq dot_x \leq max(open_j)$ **then** ▷ if trip DOT when activity open
            $b_{xj} = 1$ ▷ activity j is related to trip x

return $data[x, j, b_{xj}]$ ▷ return trip number, activity number, and whether the activity is related to the trip

Algorithm 2 Elbow method

Input: Results of K-means clustering (K from K_{min} to K_{max})
Output: K_{best} ▷ the best value K

for each K value from K_{min} to K_{max} **do**
  w=WCSS(k) ▷ Calculate total within-cluster sum of squares (WCSS)
Plot w (y-axis) against K (x-axis)
K_{best} = value of K where plot has highest gradient
return K_{best}

Fig. 3(a) plots trip distance distributions to three shopping IPOIs in New York. To account for outliers, distributions are trimmed by removing top and bottom 10%, and these trimmed distance ranges for all activities are shown in Fig. 3(b). It can be seen that trip distance distributions to different POIs with the same activity are very similar (Fig. 3(a)), while trip distances distributions to different activities vary (Fig. 3(b)). We present this as evidence that distance travelled can be used as a feature for discriminating between trip purpose activity, and we integrate this into the AIM model. In AIM, we label each trip with a number x. We then conduct K-means to group the trips into different clusters. Subsequently, we use the Elbow method (Algorithm 2) to select the best K value (the number of clusters). For each cluster, we compare the mean trip distance ($mean(d_i)$) with the maximum trip distance ($max(d_j)$) for activity j. If $mean(d_i) < max(d_j)$, then we consider j as a possible activity for all trips x in cluster i ($b_{xj} = 1$). Otherwise, ($mean(d_i) >= max(d_j)$) j will not be an activity related to trips x in cluster i ($b_{xj} = 0$). For example, from Fig. 3(b), shopping trips in New York are within 7 miles ($max(d_j) = 7$). If the average trip distance of a cluster is within 7 miles, then the journeys in the cluster will be considered possible shopping trips.

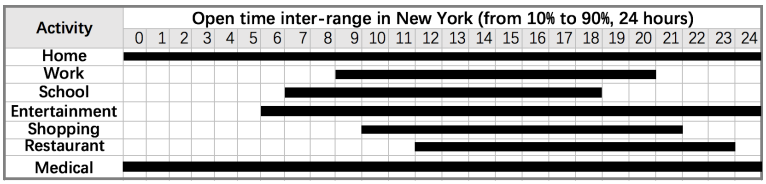


Figure 4. Opening time range (24 hours) in New York.

Table 1. Example of how to intersect the three clustering results, showing the intersection result is 'shopping'.

K-means Cluster	$Cluster_{distance}$	$Cluster_{DOP}$	$Cluster_{DOT}$	Intersection
Potential Activities	**Shop**, Medical, School	Work, **Shop**	Medical, **Shop**, Work	**Shop**

For DOP clustering (Algorithm 1, $C = DROP\_OFF\_POINT$), the number of dimensions is equal to the number of possible activities. For New York and Shenzhen, we consider eight dimensions when performing DOP clustering on activity. After we obtain the best K, we calculate the mean distance ($mean(dop_{ij})$) from the DOPs in cluster i to activity j. If $mean(dop_{ij}) < 500$ m, then we consider j to be a possible activity of all trips, x, in cluster i ($b_{xj} = 1$). Otherwise, j will not be an activity related to trips in cluster i ($b_{xj} = 0$). For example, if the mean distance from a cluster i to the nearest shopping mall j is 300 m ($mean(dop_{ij} = 300 < 500)$) then all the trips in the cluster will be labelled as possible shopping trips.

For drop-off time clustering (Algorithm 1, $C = DROP\_OFF\_TIME$), we first discover opening times (24 hours) of each activity using all POI data. Once again, opening times are calculated using the 10%-90% inter-decile range for each activity (see Fig. 4 for opening times of activities in New York). We then divide original trips into different clusters according to the time of drop-off and calculate the mean value of the drop-off time of the cluster i (mean(dot_i)) and the opening time of j (from min($open_j$) to max($open_j$)). If the cluster's mean drop-off time matches the activity's opening time ($min(open_j) \leq mean(dot_i) \leq max(open_j)$), then we iterate through all trips, x, in the cluster and if the drop-off time of the trip is within activity opening times ($min(open_j) \leq dot_x \leq max(open_j)$), j is considered a possible activity associated with the trip ($b_{xj} = 1$). For example, let i be a cluster containing trips x and y. From Fig. 4, we see opening times for shops in New York are from 10:00 to 22:00 hours. Therefore, if cluster i has mean drop-off time of 9 pm ($mean(dot_i) = 21$), then trip x with drop-off time 8 pm ($dot_x = 20$) will be considered a possible shopping trip, but trip y with drop-off time 10:30 pm ($dot_y = 22.5$) will not be considered a possible shopping trip.

After clustering, the selected activity for each taxi journey is calculated as the intersection of activities from each of the three clustering procedures. Table 1 shows a worked example for a given taxi journey, x, with suggested activities for each clustering process presented. We intersect these activities to get the final result, such that x is labelled as a shopping trip.

An examination of the three clustering results shows that some trips have multiple possible activities. To account for this, we use the Bayesian rule to set up a visit probability threshold and use the Monte Carlo method in the POI selection process to simulate individual uncertainty. The visit probability function to each POI is determined as (Gong *et al.* 2016):

Algorithm 3 Monte Carlo simulation

Input: a set of filtered trips
Output: trip purpose
for each drop off point **do**
    the visit probability to n potential POIs are $p_1, p_2, \ldots, p_n$ ▷ $\sum_{i}^{n} p_i = 1$
    set $p_0 = 0$
    generate random value, $r \in [0, 1]$
    for $i = 0$ to $n - 1$ **do** ▷ decide trip purpose (POI)
        if $\sum_{j=0}^{i} p_j \leq r < \sum_{j=0}^{i+1} p_j$ **then**
            result = POI$[i + 1]$
return result

$$Pr(O_i|(x,y),t) = \frac{Pr((x,y)|O_i,t)Pr(O_i|t)Pr(t)}{Pr((x,y),t)} \tag{2}$$

where $Pr(O_i|(x,y),t)$ represents the probability that a trip is intended for POI activity i in study area O, given that the DOP is at location (x,y) at time t. Since we have used the drop-off time clustering to filter all suitable times, we do not need to consider drop off time t, and hence the probability function becomes:

$$Pr(O_i|(x,y)) = \frac{Pr((x,y)|O_i)}{Pr(x,y)} \tag{3}$$

Subsequently, we apply a distance decay function to simulate trip purposes:

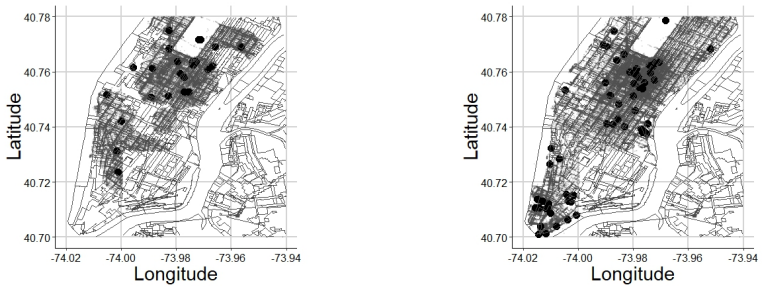
$$Pr(O_i|(x,y)) = \frac{d((x,y),O_i)^{-\beta}}{\sum_j d((x,y),O_j)^{-\beta}} \tag{4}$$

where d is the Euclidean distance from drop-off location (x,y) to all possible POI j in study area O and β is the distance decay parameter. The parameter $\beta = 1.5$ is selected to be consistent with the following studies in the existing literature: (Li *et al.* 2000), (Gao *et al.* 2013), and (Kang *et al.* 2012). Equation (4) is used to probabilistically infer the purpose of each trip. To estimate individual choice (randomness) at each DOP, we use a Monte Carlo process to simulate uncertainty such that a visit location is selected using the distribution of visit probabilities to all POIs (see Algorithm 3).

4.1. *AIM study in New York*

4.1.1. Data

We take Midtown and Lower Manhattan, New York as our study area, since it has comprehensive activities and high passenger volume. Two million taxi trips from 1 June (Monday) to 7 June (Sunday) in 2015 are used in the study (NYC-OpenData 2018). The structure of the data is shown in Table 2. The data pertaining to taxi trips are cleaned by removing invalid points caused by the following positioning errors or transfer errors: (i) delete trip data where PUPs are not in New York; (ii) delete trip data wherein trip distance is less than 500 m (0.31 miles) or more than 100 km (62 miles). It has been shown before that driving trips over 100 km are likely to involve inter-city travel, or be a result of data errors (Gong *et al.* 2016). Therefore, since we are only considering travel within one city area, we censor trips over 100 km. After cleaning, we obtain data for 2,008,752 trips in the set.



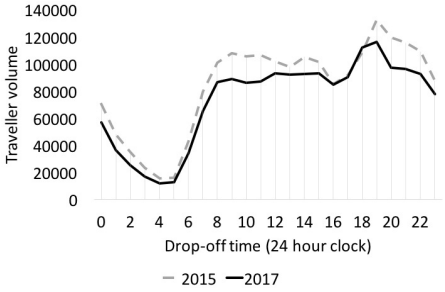
(a) Shopping activity (b) Work-related activity

Figure 6. New York: AIM results on shopping and work-related activities. Large black dots: related POIs. Small grey dots: taxi DOPs.

Table 2. Taxi data format.

Field	Data type	Example
date	int	20150601
pick up time	int	36000
drop-off time	int	79200
pick-up longitude	float	-73.822404
pick-up latitude	float	40.734424
drop-off longitude	float	-73.796814
drop-off latitude	float	40.702818
trip distance	float	5.7

Figure 5. Taxi drop-off times in 2015 and 2017.

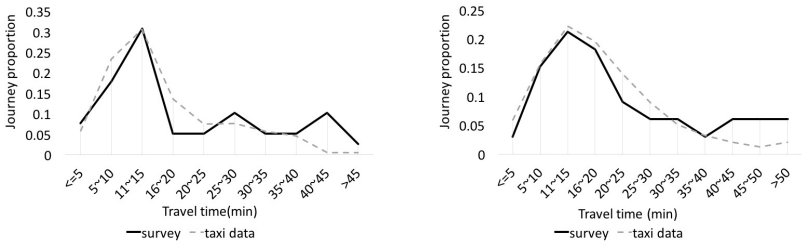


Although taxi data for New York is from 2015, only trip diaries recorded in 2017 are available. However, we find that the travellers' distribution are very similar between 2015 and 2017 (see Fig. 5); therefore, we infer that people in Manhattan exhibited similar travel behaviours between 2015 and 2017. We use 6,003 trip diaries of residents in New York in 2017 as the ground truth to validate the AIM results in New York. The following three questions in the trip diaries are related to this study: (Q1) How did you get to your destination? Answers include: walk, subway, bus, train, personal car, taxi, bicycle, and others; (Q2) What is your destination? Answers include: home, work, school, entertainment (park, recreational area), retail store/market, restaurant/bar, hospital, and others; (Q3) How long did the trip take? Answers require participants to enter travel time in minutes.

There are 2,003 POIs in Manhattan. According to the NY trip diaries, POIs are divided into eight categories: residential (home, hotel), work-related, shopping (including retail store, market), restaurant or bar, medical related (hospital and doctor office), school, entertainment (park, recreational places), and others.

4.1.2. Results and validation

We use K-means to perform clustering on trip distance, Euclidean distance from DOP to POIs, and drop-off time. The Elbow method estimation (Algorithm 2) shows that K = 6, K = 8, and K = 6 perform best on trip distance clustering, DOP clustering, and drop-off time clustering, respectively (Algorithm 1). Fig. 6 presents inferred shopping and work-related activities, where black dots in Fig. 6(a) and Fig. 6(b) represent shopping POIs and work-related POIs. We see that DOPs are often located around related POIs. Moreover, we see both shopping and work trips as POIs in Midtown Manhattan, while some POIs are related to work trips in Lower Manhattan.



(a) Shopping travel time distribution (b) Working travel time distribution

Figure 7. New York: travel time distribution observed in ground truth trip diaries (black) and inferred using AIM (grey). Percentage Error (PE) of travel time (AIM vs ground truth) is 3.36% on shopping trips (left) and 2.50% on work trips (right).

A total of 6,003 trip diaries are used to validate the results of AIM in New York. Particularly, we compare the travel time distribution of trip diaries with AIM results (including shopping and working activities). We also test the percentage error (PE) of the distributions, which is shown in Eq. 5:

$$PE = \frac{|P_{predicted} - P_{observed}|}{P_{observed}} \tag{5}$$

where $P_{predict}$ is the proportion of all taxi trips inferred as shopping trips, and $P_{observed}$ is the actual proportion of shopping trips declared in the trip diaries.

Fig. 7 presents the travel time distribution of shopping and work activities. The black lines represent the distribution presented in the trip diaries (the ground truth observation), while the grey dashed lines are the results generated from AIM. The figures reveal a close match between the black and grey lines and high accuracy of AIM performance (low value of PE). We also see that the travel time for 73% of shopping trips is within 20 minutes, while the travel time for 78% of work trips is within 35 minutes. This indicates that, on an average, people tend to travel farther to work than to shop when taking taxis; although there is considerable overlap between the distributions.

4.2. *AIM study in Shenzhen*

4.2.1. Data

We use taxi data from 24 September 2015 to 20 October 2015, which include over 10 million taxi trips. The data structure is similar to taxi data in New York, which is shown in Table 2. In this case, apply a cleaning process similar to Section 4.1.1. For validation, we use data from 712 questionnaire surveys conducted in Shenzhen in 2015. The following four questions in the survey are related to this study: (Q1) How did you travel to the shopping area? Answer options are walking, subway, car, bus, taxi, and bicycle. Responses reveal that taxi trips account for 7.3% of all the trips in questionnaires; (Q2) What is the aim of your trip? Answer options are shopping, entertainment (including parties or recreational events), and others; (Q3) How long did it take to travel to the shopping area? Answer options are below 10 minutes, 10-20 minutes, 20-30 minutes, and over 30 minutes; and (Q4) How long do you intend to stay in the shopping area? Answer options are below 1 hour, 1-2 hours, 2-4 hours, and over 4 hours.

The questionnaire surveys were conducted in five shopping areas across Shenzhen;

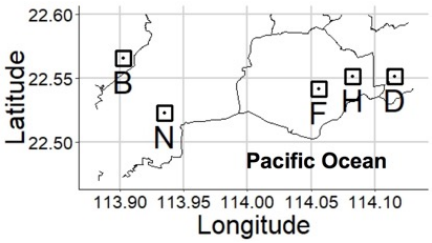


Figure 8. Map displaying the five shopping areas of interest in Shenzhen.

Table 3. Activity types in Shenzhen, based on the daily routine of individuals and the opening time of POIs. The right column lists the study areas that contain each activity type, represented by the initial letter of the area's name.

Activity Types	POIs included	located shopping area(s)
Work-Related	Office building	D,H,F,N,B
Shopping	Shopping mall, Supermarket, Cinema, Restaurant	D,H,F,N,B
Entertainment	Cultural facilities, Gym, Theatre, Park	D
Schooling	Middle School, Training institution	D,H,N,B
Medical	Hospital, Doctors' offices	D
Bank	Banks	D,H,F,N,B
Residential	House, Apartment	F,N,B
Hotel	Hotel	D,H,F,N,B

we refer to them as the 'study areas'. Fig. 8 shows the location and boundary of each study area (zones indicate boundary)—Dongmen (D), Huaqiangbei (H), Futian (F), Nanshan (N), and Baoan (B). For each study area, the POIs are located in relatively close proximity. Since 500 m is often considered as a shopping centre's influence radius (Yue *et al.* 2012), we consider a one square kilometre zone as the study area. Activities are classified into eight categories (listed in Table 3), based on people's daily routines and the opening times of POIs (Gong *et al.* 2016). Unlike New York, most restaurants in Shenzhen are located in shopping malls, and shopping behaviours often include dining activities. Therefore, we consider restaurant visits as a shopping activity. In the case study, AIM is used to identify people's shopping behaviours.

4.2.2. Results and validation

We use K-means to perform clustering on trip distance, Euclidean distance from DOP to POIs, and drop-off time. The Elbow method estimation reveals that $K = 7$, $K = 8$, and $K = 6$ perform best on trip distance clustering, Euclidean distance clustering, and drop-off time clustering, respectively. Fig. 9 presents the results of AIM in Dongmen

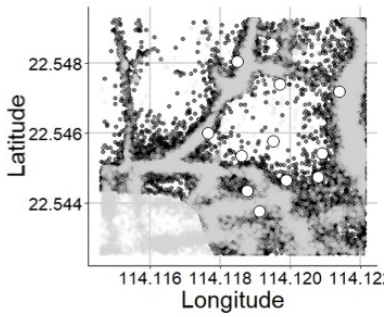


Figure 9. AIM clustering for Dongmen area, Shenzhen. Shopping POIs shown as large white circles, black dots show taxi DOPs tagged as shopping, grey dots show DOPs tagged as other activity.

Table 4. AIM output for the five shopping areas of Shenzhen. We use '-' to indicate activity is not present.

(a) Shopping behaviours (mean values of all shopping trips)

Area	time		distance (m)							
	Drop-off (pm)	Trip (min)	Shop	Work	Home	School	Hotel	Bank	Medical	Entertain
D	5	11	133.69	231.56	-	291.50	143.52	217.56	239.84	397.34
H	4	12	173.08	191.62	-	570.64	238.64	354.83	-	-
F	5	11	286.77	275.49	406.39	-	367.34	381.70	-	-
N	4	10	160.14	538.20	296.06	346.12	267.21	225.64	-	-
B	5	10	244.84	398.69	220.81	536.58	513.67	507.56	-	-

(b) Non-shopping behaviours (mean values of all non-shopping trips)

Area	time		distance (m)							
	Drop-off (am)	Trip (min)	Shop	Work	Home	School	Hotel	Bank	Medical	Entertain
D	8	19	150.61	238.59	-	306.90	157.96	210.08	244.28	418.67
H	8	18	183.85	196.98	-	561.60	240.69	365.60	-	-
F	10	15	380.58	271.70	362.60	-	426.03	350.10	-	-
N	9	17	204.05	546.78	305.99	393.73	320.72	242.68	-	-
B	9	17	409.68	320.41	273.00	411.52	349.72	521.13	-	-

Table 5. AIM validation in Shenzhen. Predicted shopping trips, as a proportion of all taxi trips, are compared with questionnaire data. Too few questionnaires were taken in Baoan for results to be statistically significant.

Area	AIM predicted shopping trips	Percentage error
Dongmen	64.99%	4.82%
Huaqiangbei	66.70%	7.60%
Futian	58.00%	0.99%
Nanshan	53.89%	0.86%
All	60.90%	5.97%

study area, wherein the black, grey, and white points represent DOPs for shopping activities, DOPs for other activities, and shopping-related POIs. From the figure, we see that most DOPs associated with shopping trips are located around shopping malls. Results (not shown) for the other four areas of study are qualitatively similar.

Table 4 shows the results of shopping and non-shopping trips. In this case, we observe the following: (i) for shopping trips, walking distance to the shopping POIs is shorter than most other POIs, as we would expect; (ii) shopping trips start later in the day (4 pm to 5 pm) than non-shopping trips (8 am to 10 am). Since shopping malls open much later (10 am) than most other POIs, this time lag is to be expected; and (iii) the average travel time of shopping trips (10 to 12 minutes, estimated from pick-up time and drop-off time) is shorter than non-shopping trips (15 to 19 minutes).

To measure the PE of the results, we use data from 712 questionnaire surveys in Shenzhen to measure the PE of the AIM prediction (Eq. 5), and thereby validate its results. The results are shown in Table 5. We see that in all four shopping areas, the percentage errors are small (right column), indicating that AIM performs well when inferring trip purpose. Since too few questionnaires were recorded in Baoan for validation to be statistically significant, we do not validate AIM in Baoan, but we do include Baoan taxi data for AIM validation across all study areas (Table 5, bottom row).

4.3. *Comparing AIM with other methods*

Table 6 compares the performance of AIM with the simple buffer radius (Yue *et al.* 2012), Furletti's Model (Furletti *et al.* 2013), and Gong's Model (Gong *et al.* 2016). We evaluate the percentage error of AIM and the buffer radius method, in Shenzhen and

Table 6. AIM evaluation and comparison with Buffer Radius, Furletti's Model, and Gong's Model. Values show PE between inferred proportion of shopping trips and ground truth proportion observed in trip diaries in New York and questionnaire data in Shenzhen. Lower values indicate better performance.

Study area	AIM	Buffer Radius	Furletti's Model	Gong's Model
NY	2.31	67.78	29.46[a]	33.93[a]
SZ	5.97	52.58		

[a]PE value is taken from the results reported in the literature using Shanghai taxi data (Gong *et al.* 2016)

New York. The values for Gong and Furletti are taken directly from results reported in the literature, using taxi data in Shanghai. We believe this to be a valid comparison since the Gong and Furletti models were validated using taxi data in a similar approach to the method we use. We see that AIM performs much better (has significantly lower PE) in both cities than the other three methods. This suggests AIM is a superior technique with general applicability across taxi GPS data sets.

The reasons AIM is superior than the others can be attributed to the following factors. First, the simple buffer radius method considers that passengers dropped-off within 500 m of the POI aim at that POI. However, when people travel to a complex environment, there are more than one POIs around a DOP, and hence it is difficult to infer activities using a buffer radius. Therefore, as expected, this method has the lowest accuracy in a complex environment with multiple POIs. Second, when compared to the Gong and Furletti models, AIM has two advantages. First, it considers trip distance when inferring activities, which connect the trip purpose (activity) with not only DOP but also PUP. With more trip information, it is likely that the AIM could have higher accuracy. Second, it uses clustering to gather data on trips that exhibit similar behaviours. Instead of inferring one trip activity using probability functions, clustering could exhibit higher performance with the help of a substantial number of trips.

5. Layer Two: Pairing Journeys Model (PJM)

Previous study shows that there is a relationship between a predecessor activity and a successor activity (Gong *et al.* 2019). In the second layer, after the AIM estimation, we develop the PJM to automatically discover outbound and return trip pairs using taxi data. In this case, we analyse shopping, medical-related, and work-related activities. The workflow of the second layer (PJM) is presented in Fig. 10. We perform two studies to evaluate the PJM: first, using only IPOIs as input for the PJM (Section 5.1); and second, comparing PJM results with the literature, using both IPOIs and AIM output as input for the PJM (Section 5.2). In the first study, we remove the need to use AIM by simplifying the problem of activity inference: we select three IPOIs and assume that all taxi trips with DOPs close to an IPOI are aimed at that POI. In Shenzhen, we select a large IKEA store (a shopping POI), a hospital (the Third Hospital, a medical-related POI), and a company (Tencent, a work-related POI).

To discover return journeys after three activities, we select all taxi trips (from September 24 2015 to October 20 in 2015) that drop-off within 500 m of the three POIs. Particularly, we use the following three rules to extract return trips: (i) the outbound and return trips occur on the same day. For journeys to hospital, we do not consider the situation wherein patients live in the hospital, (ii) the drop-off time of an outbound trip must be after the pick-up time of the return trip (at least 5 minutes),

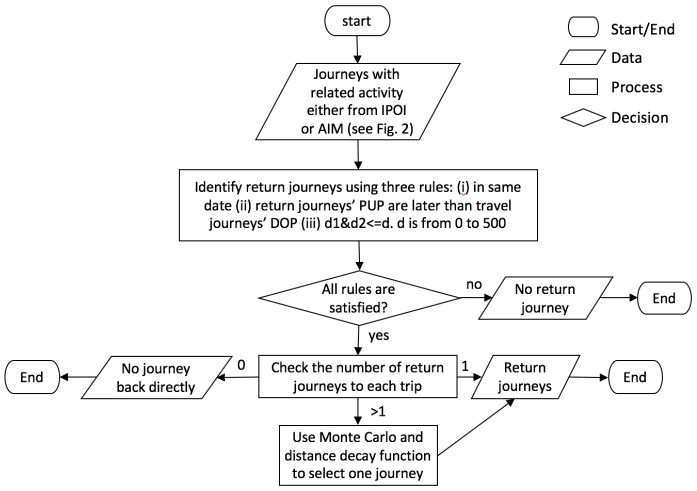


Figure 10. PJM: d_1 is the Euclidean distance between outbound DOP and return PUP; d_2 is Euclidian distance from outbound PUP and return DOP. Input taxi data has pre-tagged activities.

and (iii) d_1 refers to the Euclidean distance from the DOP of the outbound trip to the PUP of the return trip; d_2 refers to the Euclidean distance from the PUP of the outbound trip to the DOP of the return trip. When d_1 and d_2 are very small, there is a high probability that the two trips are a return journey 'pair'. The question is how to calculate the exact distance that people walk between the DOP to their destination (i.e. how to select the suitable values for d_1 and d_2). Here, we increase d from 0 to 500 m in steps of 10 m (we consider 500 m as the upper bound of the walking distance from DOP to destination, which is summarised by Yue *et al.* (2012)). The two journeys that satisfy all three rules are possible pairing journeys.

We use J_{ret} to represent the number of possible return trips to each outward trip: if $J_{ret} = 0$, then the passengers will not return after the trip (i.e. the outbound trip is non-paired); if $J_{ret} = 1$, then we can consider the trip as the return trip after the activity; if $J_{ret} > 1$, then we can use the Monte Carlo simulation to select the return trip based on d_1 and d_2 (trips with smaller d_1 and d_2 have higher probabilities of selection). Once the journey pairing is complete, we evaluate the accuracy of the pairings using the questionnaire data, which is introduced in Section 5.2.

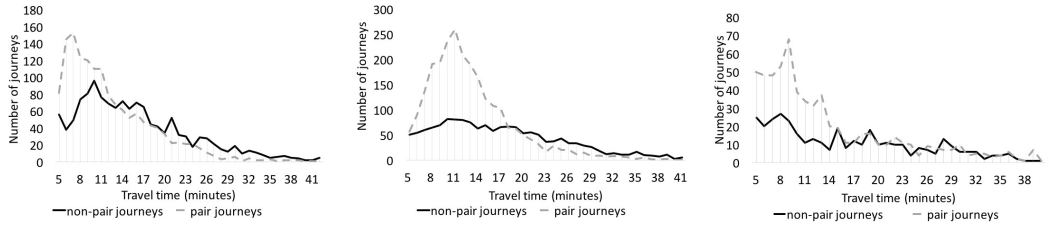
5.1. *PJM study in Shenzhen, using IPOIs*

In total, 3,075, 4,103, and 1,048 trips are collected with DOP within the 500 metre radius of IKEA, the Third Hospital, and Tencent, respectively.

We show the following three distance samples from $d_2 = 100$, $d_2 = 200$, and $d_2 = 250$ in Fig. 11: (i) when $d_2 = 100$, the resolution of the distance between the PUP in the outbound trips and DOP in return trips is similar to the distance between the opposite ends of a road. Therefore, we consider that the trips form a return pair when $d_2 \leq 100$; (ii) when $d_2 = 200$, the distance is approximately half of the width of a residential estate; additionally, we consider that the journeys form a return pair; (iii) when $d_2 = 250$, the distance is approximately from one gate to another in a residential estate in many situations (for example, the distance is similar to the distance walked from the south to the north gates). When $d_1 > 250$, we see, in some pairs of PUPs and DOPs, that the two points are not located near one POI. Therefore, we only consider possible return trips, J_{ret}, when $d_1 \leq 250$ and $d_2 \leq 250$. The results show that 55% of

(a) 100 m (b) 200 m (c) 250 m

Figure 11. Distance sample (d_2) of 'paired' journeys in Shenzhen residential area, showing outward PUP (black circle) and return DOP (grey circle).



(a) shopping journeys (b) medical journeys (c) work journeys

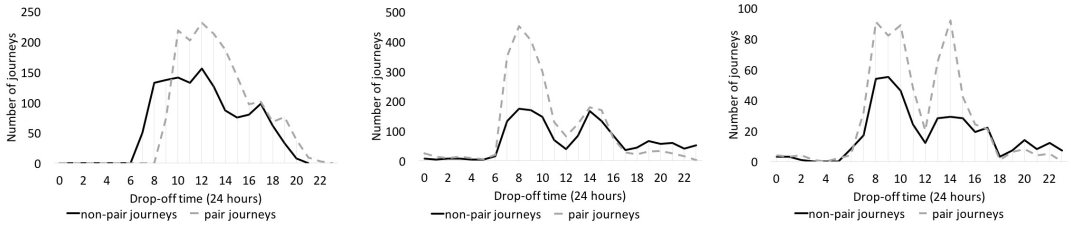
Figure 12. Shenzhen travel times to IPOIs, showing paired (grey dash) and non-paired (black line) journeys.

the people become a part of return trips after shopping, 61% of them become a part of return trips after a medical activity, and 62% of them become a part of return taxi trips after work.

Fig. 12 shows the travel distribution of shopping trips (in IKEA), medical trips (in the Third Hospital), and work trips (in Tencent). We see that trips with a short travel time are more likely to be paired. Particularly, 64% of people who travel between 5-11 minutes to shopping malls will return to the origin immediately after shopping, 70% of passengers incurring a travel time between 5-17 minutes to hospitals will return to the origin, and 71% of people who travel for less than 14 minutes to workplaces often have return trips. For other situations, the proportion of return trips roughly comprises 50% (ratio between grey (paired) dashed lines and black (non-paired) lines).

Fig. 13 presents the drop-off time distribution of different activities. During 10 am to 3 pm, 60% of people from shopping trips will return to the origin after shopping. From 1 pm to 3 pm, 70% of the passengers who travel to Tencent will have to undertake a return trip after work. During 6 to 11 am, 70% of passengers undertaking medical-related trips will return to original locations. At other times, the proportion is approximately 50%.

Fig. 14 shows the distribution of time spent on different activities (in hours). We



(a) shopping journeys (b) medical journeys (c) work journeys

Figure 13. Shenzhen drop-off times to IPOIs, showing paired (dash) and non-paired (line) journeys.

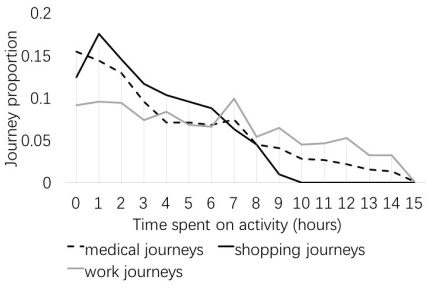
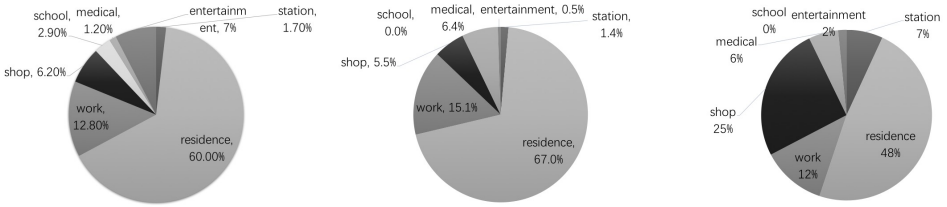


Figure 14. Shenzhen distributions of time spent on activities, calculated using PJM.



(a) shopping journeys (b) medical journeys (c) work journeys

Figure 15. PJM estimation of successor activities in Shenzhen after shopping, medical, and work activities.

see that the distribution of medical activities has a much longer tail, up to 14 hours (compared with 9 hours maximum for shopping). The proportion of people who spend less than 5 hours in shopping malls account for 81%, but 46.7% of the passengers stay more than 4 hours in hospitals. People who travel to work tend to spend 7 hours at their workplace. This difference is close to reality and meets our expectations.

We also observe the originating destinations of return trips from each POI. From Fig. 15 we see: (i) 65% and 70% of the return trips are made to residential locations after shopping and medical activities, respectively, and only 48% of the return trips are made to home after work. We also see that people would like to shop after work (25%) when compared to other predecessor activities (6.2% and 5.5% after shopping and medical treatment, respectively). Since 55% and 61% of the passengers undertaking shopping trips and medical trips, respectively, also undertake return trips, we indicate that 36% of people will return home after shopping, while 43% of people will return home after medical-related activities. (ii) while few people travel for entertainment after medical activity (0.5% when compared to 7.7% for shopping journeys and 2% for work journeys), the percentage of trips made to another hospital after a medical activity (6.7%) is higher than the other two activities (only 1.4% after shopping and 6% after work). The latter is interesting, suggesting that patients travel between hospitals after each visit. One interpretation could be hospitals offer different specialized treatments.

During the pairing journeys process, we also find an interesting phenomenon as follows: as the distance of an outbound trip increases, the d_2 decreases (that is, the farther people travel by taxi, the closer they will return to their initial PUP). To test this finding, we use the route distance to represent the taxi trip distance and Euclidean distance to estimate walking distance d_2. Min-max method is used to do normalisation before regression (Jain and Bhandare 2011). The equation of min-max normalisation

is:

$$d' = \frac{d - min(p)}{max(p) - min(p)} \tag{6}$$

where d is a value of P before normalisation, and d' is a value of d after normalisation. $min(p)$ is the minimum value of the attribute, and $max(p)$ is the maximum value of the attribute.

Fig. 16 and Table 7 show a significant negative relationship between travel distance and d_2. This is an interesting and unexpected finding. There is a strong inverse relationship between the length of the journey and the distance between the starting point of an outbound trip and the end point of a return trip. This may be attributed to the fact that a person may feel tired after a long trip and would like to return directly without walking after a drop-off. In the future, one direct application of this relationship is that researchers could use travel time and other information to estimate passengers' drop-off location when they take a return trip.

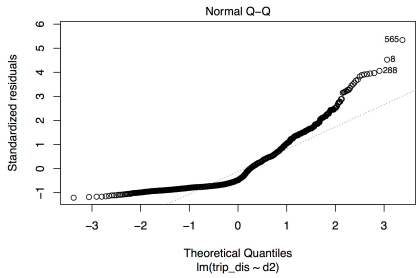


Figure 16. The Q-Q plot of travel time and d_2

Table 7. The linear regression result of estimating travel time on d_2.

Type	Value	S.E.	T value	P value
d2	-0.038	0.015	-2.594	0.009
Intercept	0.204	0.007	27.762	<2e-16

5.2. *Evaluation of PJM using AIM and IPOIs*

Here, we consider the questionnaire data in Shenzhen as ground truth to test the performance of pairing shopping trips, using IKEA paired trips and AIM paired trips (the shopping trips discovered by AIM). Previous studies have proved that different activities have different travel times (Raux *et al.* 2011, Gong *et al.* 2016). Therefore, the following two dimensions are used to validate the paired journeys process: (i) travel time and (ii) time spent on shopping.

Fig. 17 presents validation results. It is clear that each figure has a similar distribution. We also use the mean absolute percentage error (MAPE) as criteria to test the performance of the PJM (Table 8). These results provide evidence that the paired journeys process exhibits high performance with a low PE. Moreover, since paired journeys discovered from AIM have similar distribution between questionnaires and IKEA trips; therefore, we infer that AIM performs well when inferring trip purpose.

We also validate trips related to work. Agent-based simulation's results presented in literature are used to validate the work trips (Wu *et al.* 2014). The distribution of the drop-off time in pairing work trips and agent-based results are shown in Fig. 18, which shows a good match between agent-based simulation and work trips from PJM.

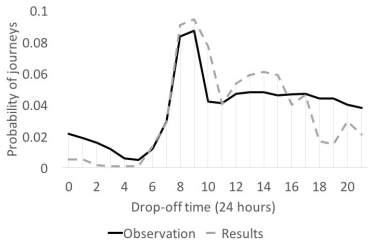


Figure 18. The drop-off time distribution between observation (ground truth) and paired journeys in work.

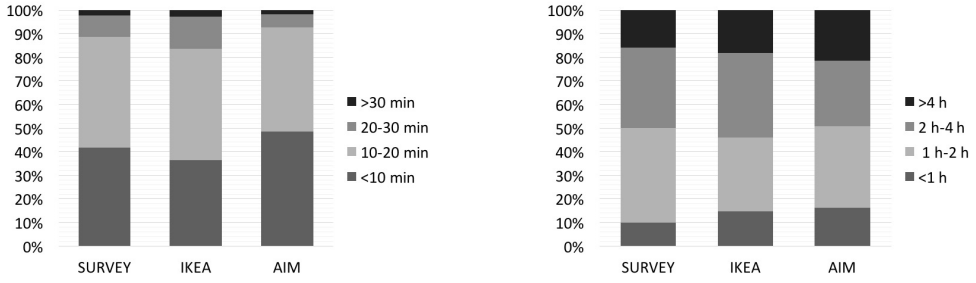
Table 8. Validation results using MAPE

Journeys	Travel time	Shopping time
IKEA	4.38%	3.27%
AIM	5.90%	5.93%

6. Conclusion

We have proposed a two-layer framework to connect people's activities with their travel routines. In the first layer, we develop the AIM to infer trip purpose. In the second layer, the PJM is proposed to identify return trips and successor activities. Results demonstrate that the AIM employs a superior method for inferring the trip purpose when compared to other existing methods in the literature. Additionally, PJM is a novel approach that makes inference possible by using activities and trip routines to estimate passengers' successor behaviours. Furthermore, results demonstrate that whether people return to their original location after a trip is related to the travel time and drop-off time. We also find that travel time has a significant negative linear relationship with the distance between the originating point of an outbound journey and the end point of a return journey.

The two-layer framework proposed in this study has a high value for many applications. For example: (i) in the medical area, the framework could use individual travel information in GPS data to infer whether the patients are satisfied with the treatment (by discovering return activities) as well as the reason they move to another medical institution. The results could be directly applied to medical online platforms, which provide pre-examination to patients; (ii) in commercial and urban planning, the framework could automatically estimate passengers' travel aims. It can determine whether they will return and where they will be dropped-off upon return. The results could provide guidance for government and companies to build infrastructure or stores in convenient locations; (iii) with an elaborate analysis, the framework could provide automatic individual travel behaviour forecasting, including information about peo-



(a) Travel time distribution

(b) Time spent on shopping

Figure 17. Pairing journeys process validation using the questionnaire data are based on the following two dimensions: travel times and time spent on shopping. The following three results are compared: questionnaires, IKEA trips, and shopping trips discovered from the AIM. The accuracy in MAPE is shown in Table 8.

ple's drop-off location, trip aim, determining the possibility and location of a return trip, and their potential residential locations. With an improvement in the accuracy of inferring trip purpose, the framework could reveal more interesting findings and developments in the future.

Acknowledgement

This research was supported by Zhejiang Natural Science Foundation (Grant No. LR17G010001), Ningbo Science and Technology Bureau (Grant No. 2017D10034, 2014A35006), UK Engineering and Physical Sciences Research Council (Grant No. EP/L015463/1), the National Science Foundation of China (No. 41671387, 91546106, 71471092), Shenzhen Scientific Research and Development Funding Program (No. CXZZS20150504141623042), and Refinitiv (formerly Thomson Reuters Financial and Risk).

References

Ashbrook, D. and Starner, T., 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7 (5), 275–286.

Beecham, R., Wood, J., and Bowerman, A., 2014. Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47, 5–15.

Furletti, B., *et al.*, 2013. Inferring human activities from GPS tracks. *In*: *ACM SIGKDD International Workshop on Urban Computing*. 1–8.

Gao, S., *et al.*, 2013. Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. *Environment and Planning B: Planning and Design*, 40 (1), 135–153.

Gong, L., *et al.*, 2016. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science*, 43 (2), 103–114.

Gong, S., *et al.*, 2019. Activity modelling using journey pairing of taxi trajectory data. *In*: *Proceedings of 4th IEEE International Conference on Big Data Analysis*, Suzhou, China. 236–240.

Gong, S., *et al.*, 2017. Geographical Huff Model Calibration using Taxi Trajectory Data. *In*: *Proceedings of the 10th ACM SIGSPATIAL Workshop on Computational Transportation Science*. ACM, 30–35.

Griva, A., *et al.*, 2016. Framing Customer Shopping Behavior Through Retail Data Analytics. *Social Science Electronic Publishing*.

Han, H., Kim, W., and Hyun, S.S., 2014. Overseas travelers' decision formation for airport-shopping behavior. *Journal of Travel & Tourism Marketing*, 31 (8), 985–1003.

Huang, L., Li, Q., and Yue, Y., 2010. Activity identification from GPS trajectories using spatial temporal POIs' attractiveness. *In*: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on location based social networks*. ACM, 27–30.

Hüttel, A., *et al.*, 2018. To purchase or not? Why consumers make economically (non-) sustainable consumption choices. *Journal of Cleaner Production*, 174, 827–836.

Jain, Y.K. and Bhandare, S.K., 2011. Min max normalization based data perturbation method for privacy protection. *International Journal of Computer & Communication Technology*, 2 (8), 45–50.

Jones, P.M., 1990. Activity analysis; State-of-the-art and future directions. *New Developments in Dynamic and Activity-Based Approaches to Travel Analysis*, 34–55.

Kang, C., *et al.*, 2012. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, 391 (4), 1702–1717.

Li, Y., Gong, S., and Liddell, H., 2000. Support vector regression and classification based

multi-view face detection and recognition. *In*: *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*. IEEE, 300–305.

Liu, Y., *et al.*, 2017. Point-of-Interest Demand Modeling with Human Mobility Patterns. *In*: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 947–955.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *In*: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif. University of California Press, 281–297.

NYC-OpenData, 2018. 2015 yellow taxi trip data. `https://data.cityofnewyork.us/Transportation/2015-Yellow-Taxi-Trip-Data/ba8s-jw6u`. Accessed June, 2018.

Pallant, J.I., *et al.*, 2017. An empirical analysis of factors that influence retail website visit types. *Journal of Retailing and Consumer Services*, 39, 62–70.

Raux, C., *et al.*, 2011. Travel and activity time allocation: An empirical comparison between eight cities in Europe. *Transport Policy*, 18 (2), 401–412.

Wang, P., *et al.*, 2017. Human Mobility Synchronization and Trip Purpose Detection with Mixture of Hawkes Processes. *In*: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 495–503.

Wu, L., *et al.*, 2014. Intra-urban human mobility and activity transition: Evidence from social media check-in data. *PloS one*, 9 (5), e97010.

Xie, K., Deng, K., and Zhou, X., 2009. From trajectories to activities: a spatio-temporal join approach. *In*: *Proceedings of the 2009 International Workshop on Location Based Social Networks*. ACM, 25–32.

Yue, Y., *et al.*, 2012. Exploratory calibration of a spatial interaction model using taxi GPS trajectories. *Computers, Environment and Urban Systems*, 36 (2), 140–153.

Yue, Y., *et al.*, 2009. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. *In*: *International Conference on Geoinformatics*. 1–6.